Reducing the Carbon Impact of Generative AI Inference (today and in 2035)

Year: 2023 | Citations: 140 | Authors: A. Chien, Liuzixuan Lin, H. Nguyen, V. Rao, Tristan Sharma

---

Abstract

Generative AI, exemplified in ChatGPT, Dall-E 2, and Stable Diffusion, are exciting new applications consuming growing quantities of computing. We study the compute, energy, and carbon impacts of generative AI inference. Using ChatGPT as an exemplar, we create a workload model and compare request direction approaches (Local, Balance, CarbonMin), assessing their power use and carbon impacts. Our workload model shows that for ChatGPT-like services, inference dominates emissions, in one year producing 25x the carbon-emissions of training GPT-3. The workload model characterizes user experience, and experiments show that carbon emissions-aware algorithms (CarbonMin) can both maintain user experience and reduce carbon emissions dramatically (35%). We also consider a future scenario (2035 workload and power grids), and show that CarbonMin can reduce emissions by 56%. In both cases, the key is intelligent direction of requests to locations with low-carbon power. Combined with hardware technology advances, CarbonMin can keep emissions increase to only 20% compared to 2022 levels for 55x greater workload. Finally we consider datacenter headroom to increase effectiveness of shifting. With headroom, CarbonMin reduces 2035 emissions by 71%.